UNITED STATES DEPARTMENT OF COMMERCE
Bureau of the Census
Washington, DC 20233-0001

MASTER FILE

October 22, 1999

CENSUS 2000 PROCEDURES AND OPERATION MEMORANDUM SERIES R-20

MEMORANDUM FOR Maureen P. Lynch
Assistant Division Chief, ICM Processing
Decennial Statistical Studies Division

From: Donna Kostanich FN for DK
Assistant Division Chief, Sampling and Estimation
Decennial Statistical Studies Division

Prepared by: Douglas Olson DO

Subject: Accuracy and Coverage Evaluation Survey – Identification and Sampling of Block Clusters for Targeted Extended Search

I. Introduction

In 1990, every block cluster sampled in the Post Enumeration Survey (PES) was also searched for persons in the surrounding blocks in an operation called "Surrounding Block Search". Experience with that operation demonstrated that searching around every cluster was not necessary because most clusters did not show any errors in the original P- or E- samples related to geography. What errors there were tended to occur in bunches, in which large groups of housing units were incorrectly enumerated or matched because of an error in locating them geographically. Because such errors tend to be clustered, it was decided for the 2000 Accuracy and Coverage Evaluation (A.C.E.) that only 20 percent of clusters would have their surrounding areas searched, and that the clusters would be selected based on a criterion that will result in the sample inclusion of a high number of clusters with large numbers of census geocoding errors. Such a search is called a "Targeted Extended Search" or "TES."

In addition to the deliberate selection of "bad" clusters, the search will be targeted by address. Before the A.C.E. person interviewing (during which TES will be performed concurrently) the Census Bureau will conduct Housing Unit Matching. This process will identify what housing units represent the Geocode Errors or Address Non-matches that flag the need to search in the surrounding area. Unlike in 1990 Surrounding Block Search, the 2000 TES will search only for the addresses that are flagged as Geocode Errors or Address Non-matches, which will make them candidates for TES. The persons living in such housing units will be considered as "TES persons" and weighted in the P- and E-samples to reflect the probability that their block cluster had been chosen for TES.

These specifications describe the steps necessary to select the TES clusters and assign them TES

weights. Those data, along with additional cluster information needed to check the validity of the sample selection, will be included in the Sample Design File.

II. Overview of Process

In outline, there are four basic steps to selecting the TES clusters:

- Determine sampling parameters and create a TES Parameter File
- Select TES Clusters
- Update the Sample Design File with the results of TES sample selection
- Verify the sample selection

Because some blocks are going to generate more P-sample and E-sample persons through TES than others, it would be desirable to design a sample in which clusters have different probabilities of being included in TES. Before the TES sample has to be selected, there will be certain information available about the clusters that will be helpful in targeting the selection of clusters for TES. Specifically, we would like to be certain to include in TES all clusters which:

- Were re-listed
- Have many housing units coded as address non-match, unresolved address match or geocode error in housing unit matching
- Have a high weighted total of housing units in these categories

Therefore, the TES clusters will be of two types:

- Those included in the TES with certainty because they show the characteristics above
- Clusters selected randomly from those that do *not* show these characteristics

The TES interview workload will include

- All re-listed clusters,
- The 5 percent of clusters with the most census geocode errors and A.C.E. address nonmatches
- The 5 percent of clusters with the most weighted census geocode errors and address nonmatches
- 10 percent of the remaining 90 percent of clusters (excluding relisted and list/enumerate clusters) selected using a systematic sample

At the end of the selection, all clusters will be assigned a variable TESFLAG that identifies their TES status and a variable TETES, a weight that reflects the probability that the cluster be included in TES. There are a few types of clusters that require special handling:

- The Puerto Rico A.C.E. operations are similar but independent of the United States. The

TES selection for the Puerto Rico A.C.E. operations will be implemented independently.

- List/Enumerate clusters will not be included in this part of TES selection. L/E clusters are out of scope for sample selection purposes. TES special procedures for these areas are currently under development.

II. Detailed Sampling Procedures

The TES sample selection is a national-level process, one of the few "one-shot" A.C.E. related operations. It will be implemented completely at once. All preliminary operations, listed under "Assumptions" below, must be completed before the TES clusters are selected.

Assumptions

Before the beginning of the TES sample selection, the following will be complete and available:

- A.C.E. sample selection and A.C.E. small block subsampling.
- A.C.E. sample reduction. The above sampling activities will be reflected on the Sample Design File.
- All initial housing unit matching operations.
- Identification of clusters to be relisted.

Large-block subsampling will *not* be completed in time for TES sample selection.

Data Sources

1. Input Files:
    a. HUMARCS_ACCT2K (MaRCs housing unit account file)-- This is a block cluster level file which includes one record for each cluster in the A.C.E. sample. This file will include the results of all the above sampling activities, as well as the initial housing unit matching results. For example, for each block cluster it will have a count of independent listing addresses not matched to census addresses that were confirmed to exist in the A.C.E. block cluster.
    b. ACE2000_SDF (Sample Design File) – include all listed clusters, whether selected for A.C.E. or not. Before starting TES, it will include the final weight of all clusters in the A.C.E. sample, including small block subsampling weights, and additional sampling related codes

2. Created during processing:
    a. TESPARAM (TES Parameter File) – Includes only two records with several variables to be used in TES sample selection.
    b. TESCLUST – file of all A.C.E. sample clusters, which will include variables and information required for TES sample selection.

3. Output File:

ACE2000_SDF?.<mmddyy> (The Sample Design File, The "?" refers to the version number of the Sample Design File, which will be updated on a flow basis. "<mmddyy>" is the date on which the most recent version of the Sample Design File was created.)-- The TES sampling operation will generate a subsequent version of the Sample Design File, with an updated version number and date. The data used in selecting the TES sample, and the sampling output itself will be included in this file for subsequent use during several production operations.

Operations

A. Create TES Parameter File (TESPARAM)

File TESPARAM (see file layout in the Attachment) contains parameters that will be used to select the TES sample. It will have two records, one for the U.S. and one for Puerto Rico. The first record has PRFLAG set equal to 0 to indicate that this record is for the U.S. The second record is for Puerto Rico. Set PRFLAG equal to 1. This record holds the sampling parameters for Puerto Rico. The other fields that need to be set before beginning the process are:

- TESRATE - The fraction (expressed as a decimal) of clusters that will be included in the TES sample. The 20 percent sampling fraction **does not** include relisted clusters. Set TESRATE equal to .20.

- UNWCRATE - The fraction (expressed as a decimal) of clusters that will be used for TES selected with certainty based on the *unweighted* number of interesting housing units (to be defined later). Set UNWCRATE to .05.

- WGTCRATE- The fraction (expressed as a decimal) of clusters that will be used for TES selected with certainty based on the *weighted* number of interesting housing units (to be defined later). Set WGTCRATE to.05.

- SUMORDIF – Flag to indicate wether TES will be based on the sum or difference of the interesting housing units in the P- and E-samples. Flag equals 1 if using the sum, 0 if using the difference.

- The remaining fields (NUMCLUST, RELISTCT, LECOUNT, UNCERNUM, WTCERNUM, SAMPSIZE and TETES) will be updated later and initially must be set to zero.

B. Create TES Cluster File (TESCLUST).

This file will contain one record for every A.C.E. sample cluster and includes the information needed to perform TES sample selection. This file will include all the records and a subset of data

fields from the HUMaRCS Account File plus additional variables from the Sample Design File.

1. For each record in the Account file copy to file TESCLUST the fields:

   CLUST - cluster number
   CURCI - housing units with match code "CI"or confirmed address non-match
   CURUI - housing units with match code "UI" or unconfirmed address match
   CURGE - housing units with match code "GE" or census geocode error
   RELIST - relist flag = 1 if cluster re-listed, 0 otherwise
   STATE - 2-digit FIPS state code
   CMDONE - Computer match done code

2. From the Sample Design File, add the following fields into TESCLUST, using the same variable names:

   WEIGHTC - A.C.E. cluster weight
   SS - A.C.E. sampling stratum
   ARST - A.C.E. sample reduction stratum
   SBCSS - small block cluster sampling stratum

3. Create additional fields that will be used in the TES selection, and assign initial values:

   Variables SAMPSTRT and PRFLAG will be used to identify sampling stratum and Puerto Rico/United States sampling process, respectively.

   SAMPSTRT, concatenate fields STATE, SS, ARST, SBCSS
   PRFLAG=1 if STATE=72, and PRFLAG=0 otherwise

   We want to target for certainty inclusion in TES clusters that have many address non-matches, unresolved address matches and census geocode errors. This information was extracted from the Housing Unit Account file in fields CURCI, CURUI and CURGE. We will need to know the total number of such housing units, and both a weighted and unweighted basis. The criterion that will be used to select the sample is a function of these counts. The exact form of the function is yet to be determined. Therefore, we have to create several variables to hold the weighted and unweighted sum and difference of these totals, and one variable to use as a sort variable once the sample selection criterion is agreed upon. Compute the following variables:

   SUMUNIHU = CURCI + CURUI + CURGE
   DIFUNIHU = Absolute value of (CURCI + CURUI- CURGE)
   SUMWTIHU = WEIGHTC * SUMUNIHU, rounded to nearest integer
   DIFWTIHU = WEIGHTC * DIFUNIHU, rounded to nearest integer
   SRTUNIHU = SUMORDIF*SUMUNIHU + (1 - SUMORDIF)*DIFUNIHU
   SRTWTIHU = SUMORDIF*SUMWTIHU + (1 - SUMORDIF)*DIFWTIHU

A few variables will ultimately be copied to the Sample Design File to identify the TES selection. For the time being, they need to be put into TESCLUST and initialized as follows:

TESELECT="Z"
TESFLAG=0
TETES=0
TESN=0
RSTES=0

4. Get a count of the number of clusters in the U.S. and Puerto Rico. Count the total number of records in TESCLUST for PRFLAG=0 and PRFLAG=1 separately. Put the total in field NUMCLUST in file TESPARAM. The first record will show the number of in-sample A.C.E. clusters in the U.S. and the second will show the number of in-sample A.C.E. clusters in Puerto Rico.

C. Identify the List/Enumerate clusters

List/Enumerate clusters will be excluded from the TES sampling operations. For these clusters, which can be identified by variable CMDONE=5, update the selection variables to reflect that they will not be part of TES:

TESELECT="O", cluster is out-of-scope for TES
TESFLAG=2, cluster is not eligible for TES
TETES=<blank>, TES Weight not relevant for these clusters
TESN=0, no order assignment used
RSTES=<blank>, random start not relevant for these clusters

D. Identify relisted clusters

All relisted clusters, identifiable by variable RELIST=1, will be included in TES. For each record in TESCLUST, if RELIST=1, set the following variables:

TESELECT="R", cluster was relisted
TESFLAG=1, cluster will be included in TES
TETES=1, the cluster's TES Weight will equal one
TESN=0, no order assignment used
RSTES=0, random start not used

E. Selection of Certainty Clusters

There are two phases to selecting the certainty cases for TES–weighted and unweighted. The clusters (records) with the highest totals of SRTUNIHU and SRTWTIHU will be

selected with certainty for inclusion in TES.

1. Since Re-list and List/Enumerate clusters are not of interest for the rest of the process, get counts of both types of assignments and put them into the TES Parameter file. For both PRFLAG=0 and PRFLAG=1, count the number of TESELECT="R" records and put the total in the TESPARAM field RELISTCT . Count the number of TESELECT="O" clusters and put the total in TESPARAM field LECOUNT.

2. Calculate the number of weighted and unweighted certainty cases needed. Using the TESPARAM file, calculate separately for PRFLAG=0 and =1:

UNCERNUM = (NUMCLUST - LECOUNT) x UNWCRATE, rounded to nearest integer.
WTCERNUM = (NUMCLUST - LECOUNT) x WGTCRATE, rounded to nearest integer.
If either of the variables UNCERNUM or WTCERNUM is negative, set to zero.

3. First, we'll select the certainty cases based on unweighted criteria. Sort TESCLUST by the variables TESELECT, PRFLAG, SRTUNIHU, WEIGHTC, CLUST from largest value to smallest value.

4. The clusters are in the order in which we want them to select the unweighted certainty cases. Only clusters that currently have TESELECT="Z" are eligible, and we need separate draws for PRFLAG=0 and PRFLAG=1. So, within the group TESELECT="Z" and PRFLAG=0, and again within TESELECT="Z" and PRFLAG=1, select the first UNCERNUM records for inclusion with certainty. For those records set the fields:

   TESELECT="U", cluster was selected with certainty based on unweighted criteria
   TESFLAG=1, cluster will be included in TES
   TETES=1, the cluster's TES Weight will equal one
   RSTES=0, randomization not used in selecting cluster
   TESN=0, cluster order not used

5. Now select the certainty cases based on weighted criteria. Sort TESCLUST by the variables TESELECT, PRFLAG, SRTWTIHU, CLUST from largest value to smallest value.

6. For both PRFLAG=0 and PRFLAG=1, select the first WTCERNUM records for which TESELECT="Z" in order by the sort above. These have been selected for inclusion in TES, based on their weighted count of interesting housing units. For those records set the fields:

   TESELECT="W", cluster was selected with certainty based on weighted criteria
   TESFLAG=1, cluster will be included in TES
   TETES=1, the cluster's TES Weight will equal one
   RSTES=0, randomization not used in selecting cluster

TESN=0, cluster order not used

F. Selection of the TES sample

We want to draw a systematic sample of an appropriate percentage of the size of the original cluster universe (excluding List/Enumerate clusters.) At this writing, we think that percentage will be 10 percent, but need to design in some flexibility in case that has to be changed, so use the values in the TES Parameter file to calculate the sample size separately for PRFLAG=0 and =1, and put the result into the TES Parameter File:

SAMPSIZE = [TESRATE x (NUMCLUST - LECOUNT)] - UNCERNUM - WTCERNUM, rounded to the nearest integer.

To get a sample of the desired size, we need a take-every that produces a sample of the correct size. Calculate the take-every and put into the TES Parameter file:

TETES =
(NUMCLUST - LECOUNT - RELISTCT - UNCERNUM - WTCERNUM) / SAMPSIZE , rounded to six decimal places.

Samples will be taken separately for each PRFLAG:

1. Sort the block clusters within each PRFLAG by TESELECT, SAMPSTRT, CLUST. All subsequent operations will be performed on clusters where TESELECT="Z".

2. Number the block clusters from 1 to N, where N is the number of block clusters with the appropriate PRFLAG. Put these indexes to variable TESN.

3. get the take-every (TETES) from the TESPARAM file.

4. Generate a sequence of numbers $TESRAND_1$, ..., $TESRAND_n$ as follows:

   a. generate a random number (RN) between 0 and 1 with 10 decimal places.

   b. Calculate a random start, RSTES, which equals RN×TETES. Round this number to six decimal places. Put into field RSTES for every cluster in the sampling universe (i.e. all clusters where TESELECT="Z", for the appropriate PRFLAG.)

   c. Let $TESRAND_1$ = RSTES.

   d. Calculate $TESRAND_J = TESRAND_{J-1}$ + TAKEVERY for J = 2, 3, ...,n, where n is the largest integer such that [RSTES + (n - 1)×TAKEVERY] ≤

N.

e. Round each $TESRAND_J$ up to the nearest integer (an integer rounds to itself).

6. Each cluster with TESN equal to one of the rounded values of $TESRAND_J$, J = 1, 2, ..., n, is in the TES sample. Set the following variables:

   TESELECT="S"
   TESFLAG=1
   TETES=TAKEVERY

7. Each cluster with TESN **not** equal to one of the rounded values of $TESRAND_J$, J = 1, 2, ..., n, is not in the sample. For these clusters set:

   TESELECT="N"
   TESFLAG=0
   TETES=0

G. All clusters have now been selected into or out of TES. Copy the variables listed on page 14 into the Sample Design File.

IV . Verification

Files TESPARAM, TESCLUST and the Sample Design File will be used for verification.

A. Verify that all relisted clusters (RELIST=1 in TESCLUST) are in the Sample Design File with TESELECT="R" and TETES = 1.

B. Verify that all List/Enumerate clusters (CMDONE=5 in the Sample Design File) are in the Sample Design File with TESELECT="O" and TETES = 0.

C. Of those clusters where TESELECT="U", identify the minimum value of CURCI+CURUI+CURGE (or difference if used in sampling). Check that all clusters whose total is greater than that value are TESELECT="U" and that clusters whose total is smaller than that are not TESELECT="U". Ignore RELIST clusters in this step. Check that the number TESELECT="U" is UNCERNUM.

D. Of those clusters where TESELECT="W", identify the minimum value of CURCI+CURUI+CURGE (or difference if used in sampling), multiplied by the cluster weight. Check that all clusters whose total is greater than that value are TESELECT="W" and that clusters whose total is smaller than that are not TESELECT="W". Ignore RELIST and TESELECT="U" clusters in this step. Check that the number of TESELECT="W" is WTCERNUM.

E. Considering only clusters whose TESELECT="S" or TESELECT="N":

1. Sort by SAMPSTRT, CLUST
2. Check that the random starts are in [0, TAKEVERY)
3. Check that indexes were assigned correctly, (TESN increment by 1 for TESELECT="S" or TESELECT="N")
4. Duplicate the selection of TESRANDj using RSTES and TAKEVERY
5. Check that the number of TESELECT="N" is about eight times TESELECT="S"

V. Testing

Since the TES selection will be performed only once for the whole country, we would like to perform a dry run before the actual sample has to be selected. The Variance estimation staff will furnish to the Coverage Measurement Processing staff a set of test files corresponding in layout to those needed for the 2000 TES:

DRPARAM (equivalent to TESPARAM)
A sample HuMARCS Account file
DRSDF1 (equivalent to ACE2000_SDF)

These files will include all the variables used for 2000 TES selection in their proper fields. For purposes of the dry run, we may change some state code to STATE=72 to simulate the effect of Puerto Rico. The output expected will be an updated file DRSDF2, corresponding in layout to the 2000 Sample Design File after TES sampling is completed.

As output, we would like to receive all the output files from the TES selection:

DRPARAM (updated during processing)
TESCLUST (created during processing)
DRSDF2 (an updated version of DRSDF1, reflecting the TES selection)

Since we intend to design the files with exactly the same variable names and layout, the program should be exactly the same as the final TES program, except for the file names. The files will be delivered for testing not later than January 14, 2000 to be completed by the software testing deadline of February 15, 2000.

The target date to complete all software development and testing is February 15, 2000. That is, the TES sample selection computer system will be ready for production on 02-15- 00.

Attachment

File Layouts

File: HUMARCS_ACCT2K (MaRCs housing unit account file, one record per cluster in A.C.E.)
Fields used for TES:

Field	Description	Width	Fields	
CLUST	Cluster Number	6	1-	6
STATE	State Code	2	424-	425
CURCI	Current HU's with Match="CI"	5	320-	324
CURUI	Current HU's with Match="UI"	5	315-	319
CURGE	Current HU's with Match="GE"	5	370-	374
RELIST	Relist Flag (0=No, 1=Yes)	1	410-	410

File: ACE2000_SDFV?.mmddyy (Sample Design File, one record per listed cluster)
Fields input for use in TES:

Field	Description	Width	Fields	
CLUST	Cluster Number	5	21-	25
WEIGHTC	Unbiased weight for A.C.E. cluster	12	334-	345
SS	Sampling Stratum	1	55-	55
ARST	A.C.E. reduction stratum	2	190-	191
SBCSS	Small block cluster sampling stratum	2	306-	307

File TESCLUST (one record for each cluster in A.C.E.)

Field	Description	Field Width	Source/Initial Value
CLUST	Cluster number	6	Acct File
CURCI	Current HU's with Match="CI"	5	Acct File
CURUI	Current HU's with Match="UI"	5	Acct File
CURGE	Current HU's with Match="CI"	5	Acct File
RELIST	Relist Flag (0=No, 1=Yes)	1	Acct File
STATE	State Code	2	Acct File
CMDONE	Computer Match Done Code	1	Acct File
WEIGHTC	A.C.E. Sampling Weight	12.6	SD File
SS	Sampling Stratum	1	SD File
ARST	A.C.E. Reduction Stratum	2	SD File
SBCSS	Small Cluster Subsampling Start.	2	SD File
SAMPSTRT	Sampling Stratum for TES	7	STATE|| SS || ARST || SBCSS
PRFLAG	Puerto Rico Flag	1	=1 if STATE=72, 0 otherwise
SUMUNIHU	Sum Unweighted Interesting HU's	5	CURCI + CURUI + CURGE
DIFUNIHU	Diff. Of Unwgt. Interesting HU's	5	| CURCI + CURUI - CURGE |
SUMWTIHU	Sum of Weighted Interesting HU's	5	WEIGHTC x SUMUNIHU
DIFWTIHU	Diff. of Weighted Interesting HU's	5	WEIGHTC x DIFUNIHU
SRTUNIHU	Sort for Unwgt. Interesting HU's	5	SUMUNIHU or DIFUNIHU
SRTWTIHU	Sort of Weighted Interesting HU's	5	SUMWTIHU or DIFWTIHU
TESELECT	TES Selection Type	1	"Z"
TESFLAG	TES Selected Flag	1	0
TETES	TES Take-every	12.6	0
RSTES	Random Start used in sampling	12.6	0
TESN	Index value used in sampling	6	0

The Sample Design File (ACE2000_SDF?.mmddyy) one record per listed cluster
This file has many fields, most unrelated to TES. In addition to other fields, the following have to be added for TES, all will be copied from TESCLUST after selection is finished:

Field	Description	Field Width	Fields
CURCI	Current HU's with Match="CI"	5	676- 680
CURUI	Current HU's with Match="UI"	5	682- 686
CURGE	Current HU's with Match="CI"	5	688- 692
TESELECT	TES Selection Type	1	694
TESFLAG	TES Selected Flag	1	696
RSTES	TES Random Start	12.6	698- 709
TETES	TES Take-every	12.6	710- 721
TESN	Index value used in sampling	6	722- 727

The possible values for TESELECT in the TESCLUST and Sample Design Files:

Code	Description	Prob of Selection	TETES
Z	Initial value; should not be present after selection is completed		
R	Re-listed cluster, must be included in TES	100 percent	1
U	Certainty selection based on unweighted criterion	100 percent	1
W	Certainty selection based on weighted criterion	100 percent	1
S	Selected by 1-in-9 sampling of non-certainty cases	11 percent	9
N	Not selected for TES	89 percent	0
O	Out of Scope for TES	0 percent	<blank>

The TES Parameter File: TESPARAM (two records, containing global variables)
Input before beginning of processing:

Field	Description	Width	Initial Value
PRFLAG	Equals 1 for Puerto Rico, 0 otherwise	1	0/1
TESRATE	Overall TES selection rate	8.6	.20
UNWCRATE	Portion selected with certainty, based on unweighted count	8.6	.05
WGTCRATE	Portion selected with certainty, based on weighted count	8.6	.05
SUMORDIFF	Flag for selection based on sum or difference (1=Sum, 0=Diff)	1	0 or 1
Created during processing:			
RELISTCT	Count of re-listed clusters	5	0
LECOUNT	Count of List/Enumerate clusters	5	0
UNCERNUM	Certainty cases based on unweighted	5	0
WTCERNUM	Certainty cases based on weighted	5	0

NUMCLUST	Number of clusters from which TES drawn	5	0
SAMPSIZE	The desired number of sampled cases	5	0
TAKEVERY	Take-Every used in sampling	12.6	0